

Docket No.: (200209753-02) 1509-393

PATENT

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Application of	:	
Takahiko KAWATANI	:	
U.S. Patent Application No. n/a	:	Group Art Unit: n/a
Filed: <i>Herewith</i>	:	Examiner: n/a

For: EVALUATING COMMONALITY OF DOCUMENTS

CLAIM OF PRIORITY

Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Dear Sir:

In accordance with the provisions of 35 U.S.C. 119, Applicant hereby claims the priority of Japanese Patent Application No. 2002-326157 filed November 8, 2002. The certified copy of it is attached.

Respectfully submitted,

LOWE HAUPTMAN GILMAN & BERNER, LLP



Henry M. Zykorie
Registration No. 27,477

1700 Diagonal Road, Suite 310
Alexandria, Virginia 22314
(703) 684-1111 AML/HMZ/iyr
Facsimile: (703) 518-5499
Date: October 29, 2003

日 本 国 特 許 庁

JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office

出 願 年 月 日

Date of Application:

2002年11月 8日

出 願 番 号

Application Number:

特願2002-326157

[ST.10/C]:

[JP2002-326157]

出 願 人

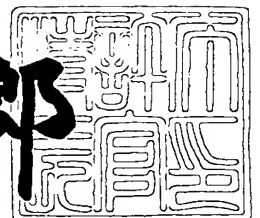
Applicant(s):

ヒューレット・パッカード・カンパニー

2003年 1月10日

特 許 庁 長 官
Commissioner,
Japan Patent Office

太田信一郎



出証番号 出証特2002-3104575

【書類名】 特許願

【整理番号】 200209753

【あて先】 特許庁長官 殿

【国際特許分類】 G06F 17/00

【発明者】

【住所又は居所】 東京都杉並区高井戸東 3 丁目 2 9 番 2 1 号 日本ヒュー
レット・パッカード株式会社内

【氏名】 川谷 隆彦

【特許出願人】

【識別番号】 398038580

【氏名又は名称】 ヒューレット・パッカード・カンパニー

【代理人】

【識別番号】 100078053

【弁理士】

【氏名又は名称】 上野 英夫

【手数料の表示】

【予納台帳番号】 061492

【納付金額】 21,000円

【提出物件の目録】

【物件名】 明細書 1

【物件名】 図面 1

【物件名】 要約書 1

【包括委任状番号】 9809395

【プルーフの要否】 要

【書類名】 明細書

【発明の名称】 文書の共通性評価方法

【特許請求の範囲】

【請求項 1】

以下の（a）から（d）のステップを有する、一つまたは複数の文書セグメントを持つ複数の文書から成る文書集合に対して、前記文書集合の各文書が話題を共通にする程度を示す文書集合共通度を求める方法、

（a）前記文書セグメント毎に、前記文書セグメントに出現する用語に対応する成分の値を 1、他の値は 0 とする文書セグメントベクトルを生成するステップと

（b）前記文書集合の各文書に対して文書セグメントベクトルより共起行列を生成するステップと、

（c）各文書の各共起行列の同一行同一列の成分の値の積により前記行前記列の成分の値を与えることによって共通共起行列を生成するステップと、

（d）前記共通共起行列の全成分、又は対角成分の和をもとに文書集合共通度を求めるステップ。

【請求項 2】

以下の（a）から（d）のステップを有する、一つまたは複数の文書セグメントを持つ複数の文書から成る文書集合に対して、各文書もしくは各文書セグメントが前記文書集合に共通する話題にどの程度近いかを示す、文書・文書集合共通度、又は文書セグメント・文書集合共通度を求める方法、

（a）前記文書セグメント毎に、前記文書セグメントに出現する用語に対応する成分の値を 1、他の値は 0 とする文書セグメントベクトルを生成するステップと

（b）前記文書集合の各文書に対して文書セグメントベクトルより共起行列を生成するステップと、

（c）各文書の各共起行列の同一行同一列の成分の値の積により前記行前記列の成分の値を与えることによって共通共起行列を生成するステップと、

（d）前記文書もしくは文書セグメントの共起行列と共通共起行列の全成分との

積和、又は対角成分の積和をもとに、文書・文書集合共通度、又は文書セグメント・文書集合共通度を求めるステップ。

【請求項 3】

以下の（a）から（e）のステップを有する、一つまたは複数の文書セグメントを持つ複数の文書から成る文書集合に対して、文書集合共通度を算出する方法、

（a）前記文書セグメント毎に、前記文書セグメントに出現する用語に対応する成分の値を 1、他の値は 0 とする文書セグメントベクトルを生成するステップと

、
（b）前記文書集合の各文書に対して文書セグメントベクトルより共起行列を生成するステップと、

（c）各文書の各共起行列の同一行同一列の成分の値の積により、但し前記同一行同一列の成分の値が 0 の場合を除く、不一致許容形の共通共起行列を生成するステップと、

（d）各文書の共起行列の各成分について値がゼロかどうかをチェックし、ゼロでない文書数を計数した共起計数行列を作成するステップと、

（e）共起計数行列の各成分について、値が所定の閾値未満の場合、前記不一致許容形の共通共起行列の対応する成分をゼロとなるように修正し、修正された前記不一致許容形の共通共起行列の全成分、又は対角成分の和をもとに不一致許容形の文書集合共通度を求めるステップ。

【請求項 4】

以下の（a）から（g）のステップを有する、一つまたは複数の文書セグメントを持つ複数の文書から成る文書集合から話題の共通する文書を抽出する方法、

（a）前記文書セグメント毎に、前記文書セグメントに出現する用語に対応する成分の値を 1、他の値は 0 とする文書セグメントベクトルを生成するステップと

、
（b）前記文書集合の各文書に対して文書セグメントベクトルより共起行列を生成するステップと、

（c）各文書の各共起行列の同一行同一列の成分の値の積により、但し前記同一行同一列の成分の値が 0 の場合を除く、不一致許容形の共通共起行列を生成する

ステップと、

(d) 各文書の共起行列の各成分について値がゼロかどうかをチェックし、ゼロでない文書数を計数した共起計数行列を作成するステップと、

(e) 共起計数行列の各成分について、値が所定の閾値未満の場合、前記不一致許容形の共通共起行列の対応する成分をゼロとなるように修正し、修正された前記不一致許容形の共通共起行列の全成分、又は対角成分の和をもとに不一致許容形の文書集合共通度を求めるステップと、

(f) 不一致許容形の文書集合共通度がある閾値以上の場合に、各文書について前記各文書の前記共起行列の全成分と前記修正された不一致許容形の共通共起行列の全成分との積和、もしくは各文書の共起行列の対角成分と前記修正された不一致許容形の共通共起行列の対角成分との積和をもとに不一致許容形文書 - 文書集合共通共通度を求めるステップと、

(g) 前記不一致許容形文書 - 文書集合共通共通度が所定の閾値を越える文書を話題が共通する文書として抽出するステップ。

【請求項 5】

前記出現する用語の種類数がMで与えられ、R個の文書からなる文書集合Dにおいて、r番目の文書を D_r 、 D_r の文書セグメント数を Y_r 、 D_r のy番目の文書セグメントベクトルを $d_{ry} = (d_{ry1}, \dots, d_{ryM})^T$ とすると、ここで、Tはベクトルの転置を表す、文書 D_r の前記共起行列 S^r は、

【数 1】

$$S^r = \sum_{y=1}^{Y_r} d_{ry} d_{ry}^T$$

で与えられることを特徴とする請求項 1 から 4 に記載の方法。

【請求項 6】

文書集合Dの共通共起行列 S^C のmn成分 S_{mn}^C は、

【数 2】

$$S_{mn}^C = \prod_{r=1}^R S_{mn}^r$$

で計算されることを特徴とする請求項 1 から 4 に記載の方法。

【請求項 7】

文書集合Dの共通共起行列 S^C の各対角成分は対応する用語の各文書の出現頻度の積によって近似することを特徴とする請求項1から4に記載の方法。

【請求項 8】

以下の（a）から（d）のステップを有する、一つまたは複数の文書セグメントを持つ複数の文書から成る文書集合に対して、前記文書集合の各文書が話題を共通にする程度を示す文書集合共通度を求める方法を情報処理装置に実現させるプログラム、

（a）前記文書セグメント毎に、前記文書セグメントに出現する用語に対応する成分の値を1、他の値は0とする文書セグメントベクトルを生成するステップと、

（b）前記文書集合の各文書に対して文書セグメントベクトルより共起行列を生成するステップと、

（c）各文書の各共起行列の同一行同一列の成分の値の積により前記行前記列の成分の値を与えることによって共通共起行列を生成するステップと、

（d）前記共通共起行列の全成分、又は対角成分の和をもとに文書集合共通度を求めるステップ。

【請求項 9】

以下の（a）から（d）のステップを有する、一つまたは複数の文書セグメントを持つ複数の文書から成る文書集合に対して、各文書もしくは各文書セグメントが前記文書集合に共通する話題にどの程度近いかを示す、文書・文書集合共通度、又は文書セグメント・文書集合共通度を求める方法を情報処理装置に実現させるプログラム、

（a）前記文書セグメント毎に、前記文書セグメントに出現する用語に対応する成分の値を1、他の値は0とする文書セグメントベクトルを生成するステップと、

（b）前記文書集合の各文書に対して文書セグメントベクトルより共起行列を生成するステップと、

（c）各文書の各共起行列の同一行同一列の成分の値の積により前記行前記列の

成分の値を与えることによって共通共起行列を生成するステップと、

(d) 前記文書もしくは文書セグメントの共起行列と共通共起行列の全成分との積和、又は対角成分の積和をもとに、文書・文書集合共通度、又は文書セグメント・文書集合共通度を求めるステップ。

【請求項 1 0】

以下の (a) から (e) のステップを有する、一つまたは複数の文書セグメントを持つ複数の文書から成る文書集合に対して、文書集合共通度を算出する方法を情報処理装置に実現させるプログラム、

(a) 前記文書セグメント毎に、前記文書セグメントに出現する用語に対応する成分の値を 1、他の値は 0 とする文書セグメントベクトルを生成するステップと、

(b) 前記文書集合の各文書に対して文書セグメントベクトルより共起行列を生成するステップと、

(c) 各文書の各共起行列の同一行同一列の成分の値の積により、但し前記同一行同一列の成分の値が 0 の場合を除く、不一致許容形の共通共起行列を生成するステップと、

(d) 各文書の共起行列の各成分について値がゼロかどうかをチェックし、ゼロでない文書数を計数した共起計数行列を作成するステップと、

(e) 共起計数行列の各成分について、値が所定の閾値未満の場合、前記不一致許容形の共通共起行列の対応する成分をゼロとなるように修正し、修正された前記不一致許容形の共通共起行列の全成分、又は対角成分の和をもとに不一致許容形の文書集合共通度を求めるステップ。

【請求項 1 1】

以下の (a) から (g) のステップを有する、一つまたは複数の文書セグメントを持つ複数の文書から成る文書集合から話題の共通する文書を抽出する方法を情報処理装置に実現させるプログラム、

(a) 前記文書セグメント毎に、前記文書セグメントに出現する用語に対応する成分の値を 1、他の値は 0 とする文書セグメントベクトルを生成するステップと、

(b) 前記文書集合の各文書に対して文書セグメントベクトルより共起行列を生成するステップと、

(c) 各文書の各共起行列の同一行同一列の成分の値の積により、但し前記同一行同一列の成分の値が 0 の場合を除く、不一致許容形の共通共起行列を生成するステップと、

(d) 各文書の共起行列の各成分について値がゼロかどうかをチェックし、ゼロでない文書数を計数した共起計数行列を作成するステップと、

(e) 共起計数行列の各成分について、値が所定の閾値未満の場合、前記不一致許容形の共通共起行列の対応する成分をゼロとなるように修正し、修正された前記不一致許容形の共通共起行列の全成分、又は対角成分の和をもとに不一致許容形の文書集合共通度を求めるステップと、

(f) 不一致許容形の文書集合共通度がある閾値以上の場合に、各文書について前記各文書の前記共起行列の全成分と前記修正された不一致許容形の共通共起行列の全成分との積和、もしくは各文書の共起行列の対角成分と前記修正された不一致許容形の共通共起行列の対角成分との積和をもとに不一致許容形文書・文書集合共通共通度を求めるステップと、

(g) 前記不一致許容形文書・文書集合共通共通度が所定の閾値を越える文書を話題が共通する文書として抽出するステップ。

【請求項 1 2】

前記出現する用語の種類数が M で与えられ、 R 個の文書からなる文書集合 D において、 r 番目の文書を D_r 、 D_r の文書セグメント数を Y_r 、 D_r の y 番目の文書セグメントベクトルを $d_{ry} = (d_{ry1}, \dots, d_{ryM})^T$ とすると、ここで、 T はベクトルの転置を表す、文書の D_r の前記共起行列を S^r が、

【数 3】

$$S^r = \sum_{y=1}^{Y_r} d_{ry} d_{ry}^T$$

で与えられることを特徴とする請求項 8 から 1 1 に記載のプログラム。

【請求項 1 3】

文書集合 D の共通共起行列 S^C の mn 成分 S_{mn}^C は、

【数 4】

$$S^C_{mn} = \prod_{r=1}^R S^r_{mn}$$

で計算されることを特徴とする請求項 8 から 11 に記載のプログラム。

【発明の詳細な説明】

【0001】

【産業上の利用分野】

本発明は文書の要約をはじめとする自然言語処理に関するものであり、特に多数の文書間の話題の共通性を定量的に評価できるようにすることによって前記処理の高性能化を図るものである。

【0002】

【従来の技術】

複数の文書からなる文書集合が与えられたとして、この文書集合の話題共通性の定量的な評価のためには次のような技術が必須である。

(A) 文書集合に共通の話題が存在するか否かを判断できるよう、各文書の話題がどの程度共通しているか数値で示す。

(B) 共通の話題に近い話題の文書、または文を文書集合から選択して全文書の共通の話題を把握できるよう、共通の話題への近さに応じて各文書、または各文にスコアを与える。

(C) 話題が全文書に共通していなくとも、話題の共通する文書群があればそれを抽出する。

【0003】

これらの内、A)については、2文書の場合には話題の共通性のスコアはその2文書間の類似度そのものと考えることができ、これまで種々の類似度の尺度が提案されてきた。最も代表的なのは余弦類似度であり、これは文書に現れる各用語の頻度を成分とするベクトルで文書を表現しておき、2文書間の類似度をそれぞれのベクトルのなす余弦によって定義するというものである。

また、B)、C)は文書集合からの共通話題の抽出に関わる技術である。このような処理は複数文書要約やTDT (Topic Detection and Tracking)などで重要な技術

となっている。従来は、共通話題の抽出は、文書のクラスタリングを行った後、クラスター毎にクラスターを代表する文や文書タイトルを選択することにより行われていた。また、最近では文やパッセージ単位にクラスタリングを行い、クラスター毎に重要なパッセージを選択することで共通話題を抽出する方法も現れている。何れにせよこれまでは共通話題の抽出にクラスタリングは欠かせない技術となっている。クラスタリングは階層的な手法と非階層的な手法に大別される。

【0004】

階層的な手法は、さらにボトムアップのアプローチとトップダウンのアプローチに分けられる。前者では、初期状態として各文書をクラスターの核とし、最も近いクラスターをマージするという処理をクラスター数が1になるまで繰り返す。これにより文書集合は木構造で表現されるようになる。後者では、全文書が1つのクラスターに属するという状態から出発し、例えばひとつのクラスター中のあらゆる文書対の中で最も低い類似度が閾値以下の場合、そのクラスターを分割するという処理を繰り返す。非階層的な手法では、予め指定された数のクラスターが何らかの基準を満たすように作成される。よく知られている方法は、ステップ1：指定されたクラスター数の文書をランダムに選択して各クラスターの中心とする、ステップ2：各文書について各クラスター中心との近さを求め、各文書を最も近いクラスターに帰属させる、ステップ3：各クラスターに帰属する文書ベクトルの平均により各クラスターの中心を求める、ステップ4：ステップ2)の処理を実行し、各文書の帰属するクラスターに変化がなければ終了、そうでなければステップ3へ、という方法である。

【0005】

【発明が解消しようとする課題】

A)については、上述のように、3文書以上の場合に対しては、2文書のときの類似度に相当する尺度は知られていなかった。そのため、同じような話題を述べている3文書の組と、4文書の組が存在したとき、“どちらの組が内容が揃っているか？”というような問題には答えようがなかった。本発明では、このような問題に対しても答えられるような尺度を提供する。

また、B)、C)の 共通話題の抽出において、ボトムアップの階層的なクラスタリ

ング処理では、各レベルのクラスターが意味のあるグルーピングとなっている保証はない。意味のあるグルーピングを指向するには、類似度が閾値を超えるクラスター対のみをマージするようにすればよいが、閾値を如何に決定するかが問題となる。トップダウンの階層的なクラスタリング処理の場合も、クラスターを分割するか否かの閾値を如何に決定するかが問題となる。また、階層的な手法では処理量の問題も無視できない。非階層的な手法では、与えられた文書集合が何個のクラスターから構成されるか事前の知識が要求されるが、これは一般的には得られない情報であり、クラスター数を正しく指定することは困難である。このようにクラスタリング技術そのものは完成された技術ではないので、共通話題の抽出を従来のクラスタリング技術を用いて行っても最適であるという保証はなかった。このようなことから本発明では従来のクラスタリング技術に依らない共通話題抽出方法を提供する。

【 0 0 0 6 】

【課題を解決するための手段】

本発明において、A)に関する課題を解決するための基本的な考え方は、文書間の情報共通量を求め、次いで求められた情報共通量を文書の長さや文書数に依存しないように正規化を行うというものである。従って、文書間の情報共通量を如何に定義して如何に求めるかが重要となるが、本発明では以下のように行っている。先ず、2つの文を考えると、2つの文の間の情報共通量は共通する用語の数で決まると考える。また、2つの文書間の情報共通量は、各文書から文を1つずつ取り出して組み合わせたとして、組み合わされた文の対における共通用語数の全組み合わせに対する和、もしくは2乗和で決まるとする。この場合文の組み合わせは各文書の文数の積通り存在することになる。3文書以上の場合も、文書間の全ての文の組み合わせを考えればよい。このような文の組み合わせにおける共通用語数の算出を容易にするため、本発明では、各文を各成分が対応する用語の有無を表す2値ベクトルで表したうえで、各文書を文ベクトルの集合で表す。また、2つ以上の文ベクトルの組み合わせに対して、共通ベクトルの概念を導入する。2つのベクトル $a=(a_n)$ 、 $b=(b_n)$ の共通ベクトルを $c=(c_n)$ とする時、本発明の場合、文ベクトルは2値なので、共通ベクトルの成分は $C_n = a_n \times b_n$ によって求めるこ

とができる。例えばベクトル(0, 1, 1, 0)と(1, 1, 0, 1)との共通ベクトルは(0, 1, 0, 0)となる。3個以上のベクトルの共通ベクトルの成分は、対応する成分同士の積となる。

【0007】

簡単な例として、6個の用語が出現し、それぞれが4、3、3個の文からなる文書 D_1 、 D_2 、 D_3 を考える。

【0008】

【表1】

文書	文	文ベクトルの成分					
D_1	D_{11}	0	1	1	0	1	1
	D_{12}	1	1	0	0	0	1
	D_{13}	1	1	0	0	1	1
	D_{14}	1	0	1	0	1	0
D_2	D_{21}	0	0	1	1	0	1
	D_{22}	1	0	1	0	1	1
	D_{23}	0	0	0	1	1	0
D_3	D_{31}	1	0	1	1	1	1
	D_{32}	0	1	1	1	0	0
	D_{33}	1	0	0	1	1	1

【0009】

文書 D_r ($r=1,2,3$)の y 番目の文を D_{ry} で表すこととする。表1はそのような文書 D_1 、 D_2 、 D_3 の文ベクトルの例を示している。表1では文書 D_r ($r=1,2,3$)の y 番目の文を D_{ry} で表している。表1の文書 D_1 、 D_2 、 D_3 の文の組み合わせは $4 \times 3 \times 3 = 36$ 通り存在することになるが、表2はそのうちの6通りについて共通ベクトルと共通用語数を示している。

【0010】

【表 2】

文の組み合わせ	共通ベクトルの成分						共通単語数
D_{11} D_{21} D_{31}	0	0	1	0	0	1	2
D_{11} D_{21} D_{32}	0	0	1	0	0	0	1
D_{11} D_{21} D_{33}	0	0	0	0	0	1	1
D_{11} D_{22} D_{31}	0	0	1	0	1	1	3
D_{11} D_{22} D_{32}	0	0	1	0	0	0	1
D_{11} D_{22} D_{33}	0	0	0	0	1	1	2
⋮	⋮						⋮
⋮	⋮						⋮

【 0 0 1 1 】

文 D_{11} 、 D_{21} 、 D_{31} の組み合わせの場合、3文書とも1となる文ベクトルの成分は、3番目と6番目であり、共通ベクトルは3番目と6番目のみが値1をとるベクトルとなる。文 D_{11} 、 D_{21} 、 D_{31} の共通用語数は共通ベクトルで値が1の成分数であるから、2となる。文 D_{11} 、 D_{21} 、 D_{32} の組み合わせの場合には、共通ベクトルは3番目の成分のみが値1となり、共通用語数は1となる。文書 D_1 、 D_2 、 D_3 の情報共通量は、36個の文の組み合わせの各々における共通用語数の和、もしくは共通用語数の2乗和である。

また本発明では、共通用語数の和、もしくは2乗和の算出を容易にするため、共通ベクトルの共起行列の概念を導入する。共通ベクトルの共起行列を S^C とすると、その成分 S^C_{mn} は各共通ベクトルの m 番目の成分と n 番目の成分との積を求め、その積の値をすべての共通ベクトルについて合計したものである。上記の例では36個の共通ベクトルを用いて S^C を求めることになる。共通ベクトルの共起行列を用いると、共通用語数の和は共通ベクトルの共起行列の対角成分の和で、共通用語数の2乗和は共通ベクトルの共起行列の全成分の和で与えられる。従って、共通ベクトルの共起行列を如何に効率的に求めるかが重要と成るが、本発明では共通ベクトルを得ることなく求める方法を提供する。

【 0 0 1 2 】

また、B)における課題を解決するためのアプローチとしては以下の2通りの考えられる。ひとつは、対象となる文書もしくは文を本来の文書集合に加えて新し

い文書集合を作成し、新しい文書集合での情報共通量を求めると、本来の文書集合の共通の話題に近い文書・文ほど上記情報共通量の値は大きくなるであろうという考え方である。2番目は、対象となる文書もしくは文と本来の文書集合から求められる共通ベクトル集合との間で類似度を求めると、この類似度の高い文書・文ほど本来の文書集合の共通の話題に近いであろうという考え方である。

【 0 0 1 3 】

C) は話題が全文書に共通せず、部分的に共通性が存在する場合を対象にしている。C) における課題を解決するためのアプローチは次の通りである。上記では、共通ベクトルは、組み合わされた文ベクトル群において全文書が値1となる成分に限って値1を与えていた。言わば全文書一致形の共通ベクトルであった。それに対して、ここでは特定の成分に着目したとき、その成分の値が1となる文ベクトルの数がある閾値を越えたときに共通ベクトルの当該成分に値1を与えるようにする。これは不一致許容形の共通ベクトルとも呼ぶべきものである。このように得られた共通ベクトル集合を用いて上記B)のアプローチを採用すれば、閾値を適当に設定することにより、部分的に存在する共通話題に対する各文書・文の近さが求められる。

【 0 0 1 4 】

上述のように、本発明によれば複数の文書の話題がどの程度共通するかをスコアで示すことができるようになり、これは文書の話題共通性の解析の重要な基本技術となる。また、全文書で話題が一致していなくとも、(1)話題を同じにする文書が含まれていればそれらを抽出し、(2)抽出された文書の話題の共通の程度のスコアを求め、(3)抽出された文書が共有する話題が端的にユーザに分かるよう共通話題に最も近い文を抽出する、という一連の処理が可能となる。これらのうち(1)(3)は従来技術によっても可能な処理であるが、本発明では各文書の各文の間の共通ベクトルという新しい概念を用いた処理がベースになっており、従来に比べより適確な結果が期待できる。

【 0 0 1 5 】

【実施例】

図1は、本発明の概要を示すブロック図である。110は文書入力ブロック、1

20は文書前処理ブロック、130は文書情報処理ブロック、140は出力ブロックを示す。文書入力ブロック110には、処理したい文書、文、文書セグメント等が入力される。文書前処理ブロック120では、入力された文書の用語検出、形態素解析、文書セグメント区分け等が行われる。文書セグメントについて説明する。文書セグメントは文書を構成する要素であり、その最も基本的な単位は文である。英文の場合、文はピリオドで終わり、その後ろにスペースが続くので文の切出しは容易に行うことができる。その他の文書セグメントへの区分け法としては、ひとつの文が複文からなる場合主節と従属節に分けておく方法、用語の数がほぼ同じになるように複数の文をまとめて文書セグメントとする方法、文書の先頭から含まれる用語の数が同じになるように文とは関係なく区分けする方法などがある。文書情報処理ブロック130は以下に詳細に説明するが、情報処理を行い、文書集合共通度、文書・文書集合共通度、文書セグメント・文書集合共通度を求めたり、共通の話題に近い文書、文書セグメント等を抽出する。出力ブロック140は文書情報処理ブロック130で得られた結果を、ディスプレイ等の出力装置に出力する。

【0016】

図3は与えられた文書集合に対して、各文書の話題がどの程度共通しているかを示す文書集合共通度を算出し、共通の話題への近さに応じて各文書、または各文書セグメントにスコアを与える本発明の第1の実施例を示す。この発明の方法は、汎用コンピュータ上でこの発明を組み込んだプログラムを走らせることによって実施することができる。図3は、そのようなプログラムを走らせている状態でのコンピュータのフローチャートである。31は文書集合入力、ブロック32は用語検出、ブロック33は形態素解析、ブロック34は文書セグメント区分けである。ブロック35は文書セグメントベクトル作成、ブロック36は文書毎の共起行列算出、37は共通共起行列算出、38は文書集合共通度算出、39は文書（文書セグメント）・文書集合共通度算出である。以下、英文文書を例に実施例を説明する。

【0017】

まず、文書集合入力31において対象となる文書集合が入力される。用語検出32において、各入力文書から単語、数式、記号系列などを検出する。ここでは、単

語や記号系列などを総称して全て用語と呼ぶ。英文の場合、用語同士を分けて書く正書法が確立しているので用語の検出は容易である。次に、形態素解析33は、各入力文書に対して用語の品詞付けなどの形態素解析を行う。次に文書セグメント区分け34において各入力文書に対して文書セグメントへの区分けを行う。文書セグメントベクトル作成35は、先ず文書全体に出現する用語から作成すべきベクトルの次元数および各次元と各用語との対応を決定する。この際に出現する全ての用語の種類にベクトルの成分を対応させなければならないということではなく、品詞付け処理の結果を用い、例えば名詞と動詞と判定された用語のみを用いてベクトルを作成するようにしてもよい。次いで、各文書セグメントに出現する用語に対応する成分のみが値1、他は0となるような文書セグメントベクトルを作成する。

【0018】

文書毎の共起行列算出36では、各文書で用語の出現頻度、用語間の共起頻度を反映するような共起行列を作成する。以降、文を文書セグメントとした場合について説明を続ける。ここでは、現れる用語集合が $\{w_1, \dots, w_M\}$ で与えられ、R個の文書から成る集合Dを考える。さらに、r番目の文書を D_r とすると、 D_r は Y_r 個の文からなるものとし、y番目の文及びその文ベクトルを D_{ry} 、 $d_{ry} = (d_{ry1}, \dots, d_{ryM})^T$ とする。ここで、Tはベクトルの転置を表す。 d_{ry} は2値ベクトルであり、 d_{rym} はm番目の用語の有無を表す。文書の D_r の共起行列を S^r とすると、これは

【0019】

【数5】

$$S^r = \sum_{y=1}^{Y_r} d_{ry} d_{ry}^T$$

・・・ (1)

【0020】

で与えられる。式(1)から分かるように、 S^r のmn成分は

$$S^r_{mn} = \sum_{y=1}^{Y_r} d_{rym} d_{ryn}$$

により与えられる。従って、 S^r_{mm} は文書 D_r において用語mが生起する文の数、 S^r_{mn} は用語mとnとが共起する文の数を表すことになる。もし同じ用語が同じ文に2

回以上出現しないのであれば、 S^r_{mm} は文書 D_r における用語 m の出現頻度となる。共通共起行列算出37では共通ベクトルを対象に共起行列 S^C を求める。これを共通共起行列と呼ぶ。前述のように、各文書から文ベクトルを1つずつ取り出して組み合わせた場合の共通ベクトルの各成分の値は各文ベクトルの対応する成分の積で与えられる。

本実施例の場合、文ベクトルはバイナリなので、共通ベクトルの成分は $C_n = a_n \times b_n$ によって求めることができる。例えばベクトル(0, 1, 1, 0)と(1, 1, 0, 1)との共通ベクトルは(0, 1, 0, 0)となる。3個以上のベクトルの共通ベクトル成分は、対応する成分同士の積となる。ここで、説明を簡単にする為に、3つの文書、 D_1 、 D_2 、 D_3 間の全ての文の組み合わせに対して求められる $Y_1 \times Y_2 \times Y_3$ 通りの共通文ベクトルの共起行列 S^C を求める。 D_1 、 D_2 、 D_3 のそれぞれの i 、 j 、 k 番目のベクトル d_{1i} 、 d_{2j} 、 d_{3k} の共通文ベクトルを $c^{ijk}_m = (c^{ij}_m, c^{jk}_m, c^{ki}_m)$ で表すと、前述のように、 c^{ijk}_m は $c^{ijk}_m = d_{1im}d_{2jm}d_{3km}$ で求められる。 S^C の各成分は

$$[0 \ 0 \ 2 \ 1]$$

【数 6】

$$\begin{aligned} S^C_{mn} &= \sum_{i=1}^{Y_1} \sum_{j=1}^{Y_2} \sum_{k=1}^{Y_3} c^{ijk}_m c^{ijk}_n \\ &= \sum_{i=1}^{Y_1} \sum_{j=1}^{Y_2} \sum_{k=1}^{Y_3} d_{1im} d_{1in} d_{2jm} d_{2jn} d_{3km} d_{3kn} \\ &= S^1_{mn} S^2_{mn} S^3_{mn} \end{aligned}$$

$$[0 \ 0 \ 2 \ 2]$$

により与えられる。さらに一般化して説明を続ける。 R 文書の場合、文の組み合わせにおいて文書 D_r から $k(r)$ 番目の文が取り出されたとして、共通ベクトルを

$$c^{k(1) \ k(2) \ \dots \ k(R)} = (c^{k(1) \ k(2) \ \dots \ k(R)}_1, \dots, c^{k(1) \ k(2) \ \dots \ k(R)}_M)$$

と書くと、

$$c^{k(1) \ k(2) \ \dots \ k(R)}_m \text{ は } d_{1k(1)m} d_{2k(2)m} \dots d_{mk(m)m}$$

と表わすことが出来るので、 S^C のmn成分は次の式で与えられる。

【 0 0 2 3 】

【数 7】

$$\begin{aligned}
 S^C_{mn} &= \sum_{k(1)=1}^{Y_1} \sum_{k(2)=1}^{Y_2} \cdots \sum_{k(R)=1}^{Y_R} c^{k(1)k(2)\cdots k(R)}_m c^{k(1)k(2)\cdots k(R)}_n \\
 &= \sum_{k(1)=1}^{Y_1} \sum_{k(2)=1}^{Y_2} \cdots \sum_{k(R)=1}^{Y_R} (d_{1k(1)m} d_{2k(2)m} \cdots d_{Rk(R)m}) (d_{1k(1)n} d_{2k(2)n} \cdots d_{Rk(R)n}) \\
 &= \sum_{i_1=1}^{Y_1} d_{1k(1)m} d_{1k(1)n} \sum_{j_1=1}^{Y_2} d_{2k(2)m} d_{2k(2)n} \cdots \sum_{k=1}^{Y_R} d_{Rk(R)m} d_{Rk(R)n} \\
 &= \prod_{r=1}^R S^r_{mn} \\
 &\quad \cdot \cdot \cdot (2)
 \end{aligned}$$

【 0 0 2 4 】

式（2）は共通共起行列の各成分は各文書の共起行列の対応する成分同士の積として求められることを示しており、共通共起行列は共通ベクトルを実際に求めることなく得ることができる。また、前述のように、同じ用語が同じ文に2回以上出現しないのであれば、 S^r_{mm} は文書 D_r における用語 m の出現頻度となる。同じ用語が同じ文に2回以上出現する頻度は少ないと考えられるので、 S^C の各対角成分は対応する用語の各文書の出現頻度の積によって近似することもできる。

文書集合共通度算出37では、各文書の話題がどの程度共通しているかを示すスコアを算出する。前述のように、本発明では各共通ベクトルで値が1の成分数の全共通ベクトルに対する和、もしくは2乗和をもとに各文書の文書集合共通度を求める。前者を線形モデル、後者を2次モデルと呼ぶ。先ず前者の線形モデルの場合について述べる。各共通ベクトルで値が1の成分数の和を $G_1(D_1, \dots, D_R)$ とする。これは、

【 0 0 2 5 】

【数 8】

$$\begin{aligned}
 G_1(D_1, \dots, D_R) &= \sum_{k(1)=1}^{Y_1} \sum_{k(2)=1}^{Y_2} \cdots \sum_{k(R)=1}^{Y_R} \sum_{m=1}^M c^{k(1)k(2)\cdots k(R)}_m \\
 &= \sum_{k(1)=1}^{Y_1} \sum_{k(2)=1}^{Y_2} \cdots \sum_{k(R)=1}^{Y_R} \sum_{m=1}^M (c^{k(1)k(2)\cdots k(R)}_m)^2 \\
 &= \sum_{m=1}^M S^C_{mm}
 \end{aligned}$$

・・・ (3)

【0026】

のように求めることができ、 $G_I(D_1, \dots, D_R)$ は共通共起行列の対角成分の和で表される。式(3)は文書集合における各文書の情報共通量を表すが、情報共通量の値は文書の長さや文書数に依存した値になるので、これらの影響を受けないように以下のように正規化し、文書集合共通度 $com_I(D)$ とする。

【0027】

【数9】

$$com_I(D) = \left[\frac{G_I(D_1, \dots, D_R)}{\sqrt[R]{G_I(D_1, \dots, D_1)G_I(D_2, \dots, D_2) \cdots G_I(D_R, \dots, D_R)}} \right]^{1/(R-1)}$$

$$= \left[\frac{\sum_{m=1}^M S_{mm}^C}{\sqrt[R]{\prod_{r=1}^R \sum_{m=1}^M (S_{mm}^r)^R}} \right]^{1/(R-1)}$$

・・・ (4)

2次モデルについて述べる。各共通ベクトルで値が1の成分数の2乗和を $G_S(D_1, \dots, D_R)$ とする。これは、

【0028】

【数10】

$$G_S(D_1, \dots, D_R)$$

$$= \sum_{k(1)=1}^{Y_1} \sum_{k(2)=1}^{Y_2} \cdots \sum_{k(R)=1}^{Y_R} \left(c^{k(1)k(2) \cdots k(R)}_1 + \cdots + c^{k(1)k(2) \cdots k(R)}_M \right)^2$$

$$= \sum_{m=1}^M \sum_{n=1}^M \sum_{k(1)=1}^{Y_1} \sum_{k(2)=1}^{Y_2} \cdots \sum_{k(R)=1}^{Y_R} \left(c^{k(1)k(2) \cdots k(R)}_m c^{k(1)k(2) \cdots k(R)}_n \right)$$

$$= \sum_{m=1}^M \sum_{n=1}^M S_{mn}^C$$

・・・ (5)

のように、共通共起行列の各成分の和で求められる。2次モデルの場合の文書集合共通度を $com_S(D)$ とすると、これは以下のように求めることができる。

【0029】

【数11】

$$com_S(D) = \left[\frac{G_S(D_1, \dots, D_R)}{\sqrt[R]{G_S(D_1, \dots, D_1)G_S(D_2, \dots, D_2) \cdots G_S(D_R, \dots, D_R)}} \right]^{1/(R-1)}$$

$$= \left[\frac{\sum_{m=1}^M \sum_{n=1}^M S_{mn}^C}{\sqrt[R]{\prod_{r=1}^R \sum_{m=1}^M \sum_{n=1}^M (S_{mn}^r)^R}} \right]^{1/(R-1)}$$

・・・ (6)

【0030】

文書（文書セグメント）・文書集合共通度算出39では、対象とする文書または文書をPとして、Pが文書集合Dの共通の話題にどれだけ近いかを示す尺度として、文書・文書集合共通度を求める。これには、次の2つの方法が存在する。

第1の方法は、Pを文書集合Dに加えた新しい文書集合の文書集合共通度を文書・文書集合共通度とする方法である。文書Pの共起行列を S^P として、線形モデル、2次モデルの場合の文書・文書集合共通度をそれぞれ $com_1(D+P)$ 、 $com_s(D+P)$ とすると、これらは以下のように求めることができる。

【0031】

【数12】

$$com_1(D+P) = \left[\frac{\sum_{m=1}^M S_{mm}^C S_{mm}^P}{\sqrt{R+1} \left(\sum_{m=1}^M (S_{mm}^P)^{R+1} \right) \prod_{r=1}^R \sum_{m=1}^M (S_{mm}^r)^{R+1}} \right]^{1/R} \dots (7)$$

【0032】

【数13】

$$com_s(D+P) = \left[\frac{\sum_{m=1}^M \sum_{n=1}^M S_{mn}^C S_{mn}^P}{\sqrt{R+1} \left(\sum_{m=1}^M \sum_{n=1}^M (S_{mn}^P)^{R+1} \right) \prod_{r=1}^R \sum_{m=1}^M \sum_{n=1}^M (S_{mn}^r)^{R+1}} \right]^{1/R} \dots (8)$$

【0033】

第2の方法は、Pから求められる共起行列と共通共起行列との類似度により文書・文書集合共通度を定義する方法である。これには共起行列の対角成分のみを用いる場合と全成分を用いる場合の2通りが考えられる。文書・文書集合共通度を前者について $com_1(D, P)$ 、後者について $com_s(D, P)$ と表記すると、

【0034】

【数14】

$$com_1(D, P) = \frac{\sum_{m=1}^M S_{mm}^C S_{mm}^P}{\sqrt{\sum_{m=1}^M (S_{mm}^C)^2} \sqrt{\sum_{m=1}^M (S_{mm}^P)^2}} \dots (9)$$

【 0 0 3 5 】

【数 1 5】

$$com_2(D, P) = \frac{\sum_{m=1}^M \sum_{n=1}^M S_{mn}^C S_{mn}^P}{\sqrt{\sum_{m=1}^M \sum_{n=1}^M (S_{mn}^C)^2} \sqrt{\sum_{m=1}^M \sum_{n=1}^M (S_{mn}^P)^2}}$$

・・・ (10)

【 0 0 3 6 】

によって求めることができる。第1の方法、第2の方法とも対象とする文書もしくは文の共起行列と共通共起行列の全成分もしくは対角成分の積和をもとに求められる。

図4は、話題が必ずしも共通しない文書集合から話題の共通する文書群を抽出する本発明の第2の実施例を示す。この発明の方法は、汎用コンピュータ上でこの発明を組み込んだプログラムを走らせることによって実施することができる。図4は、そのようなプログラムを走らせている状態でのコンピュータのフローチャートである。31は文書集合入力、ブロック32は用語検出、ブロック33は形態素解析、ブロック34は文書セグメント区分けである。ブロック35は文書セグメントベクトル作成、ブロック36は文書毎の共起行列算出、ブロック47は不一致許容形共通共起行列及び共起計数行列の算出、ブロック48は不一致許容閾値設定、ブロック49は不一致許容形文書集合共通度算出、ブロック50は不一致許容形文書・文書集合共通度算出、及び文書選択、ブロック51は選択された文書集合の文書集合共通度算出、及び妥当性判定、ブロック52は不一致許容閾値変更である。これらのうち、31～36は図3に示したものと全く同じである。

【 0 0 3 7 】

図3の場合と同じように文が文書セグメントとなっていることを想定する。不一致許容形共通共起行列及び共起計数行列の算出47における不一致許容形共通共起行列の各成分の算出では、各文書の共起行列の内その成分の値がゼロでない共起行列のみが用いられる。従って、ある用語、用語共起が文書集合Dに必ず現れる限り、不一致許容形共通共起行列の対応成分は0以外の値をとる。このような行列をTと表記する。さらに、47では各用語、または各用語対の生起、または共起した回数を保持する行列Uを求める。行列TとUは、図2に示されるように以下の

ような手順で求めることができる。

ステップ 6 1 $r=1$ とおく。T の全成分は 1、U のそれは 0 とする。

ステップ 6 2. $S_{mn}^r > 0$ のとき、

$$T_{mn} = S_{mn}^r T_{mn}$$

$$U_{mn} = U_{mn} + 1 \quad (\text{ステップ 6 3})$$

ステップ 6 4. $r=R$ で終了。そうでなければ $r=r+1$ (ステップ 6 5) としてステップ 6 2 へ行く。

【 0 0 3 8 】

不一致許容閾値設定 48 では後段の処理のために閾値 A の初期値を設定する。閾値 A は不一致許容形共通共起行列 T において A 個以上の文書で現れる用語、もしくは用語共起のみを有効にするために用いられる。閾値 A の初期値は共起計数行列 U の各成分の中での最大値である。

ブロック 49 では、A 個以上の文書で現れる用語、もしくは用語共起に対応する成分以外は値を 0 とした不一致許容形共通共起行列を用いて不一致許容形の文書集合共通度を算出し、閾値処理を行ってブロック 50 の処理に移行するか否かの判断を行う。上記のように修正された不一致許容形共通共起行列を T^A とすると、行列 T^A の mn 成分は以下のように決められる。

$$T_{mn}^A = T_{mn}, \text{ if } U_{mn} \geq A,$$

$$T_{mn}^A = 0 \quad \text{otherwise.}$$

図 3 の場合は共通ベクトルにおいて値が 1 となる成分は文の組み合わせにおいて全ての文ベクトルで値が 1 となる成分のみであったが、ここでは、A 文書以上で現れる用語に対応する成分が値 1 となるように共通ベクトルを決定したことになる。

行列 T^A は、そのように決定された全共通ベクトルから求められる共通共起行列である。式 (4) 式 (6) において行列 S^C の替りに行列 T^A を用いた文書集合共通度を不一致許容形の文書集合共通度と呼ぶこととして、線形モデルの場合は

【 0 0 3 9 】

【 数 1 6 】

$$com_l(D; T^A) = \left[\frac{\sum_{m=1}^M T_{mm}^A}{\sqrt[R]{\prod_{r=1}^R \sum_{m=1}^M (S_{mm}^r)^R}} \right]^{1/(R-1)}$$

．．． (11)

により、2次モデルの場合は

【0040】

【数17】

$$com_s(D; T^A) = \left[\frac{\sum_{m=1}^M \sum_{n=1}^M T^A_{mn}}{\sqrt[R]{\prod_{r=1}^R \sum_{m=1}^M \sum_{n=1}^M (S^r_{mn})^R}} \right]^{1/(R-1)}$$

．．． (12)

【0041】

のように求める。 $com_1(D; T^A)$ 、 $com_s(D; T^A)$ は行列 T^A を共通共起行列として用いて求められる文書集合共通度という意味である。 $com_1(D; T^R)$ 、 $com_s(D; T^R)$ は $com_1(D)$ 、 $com_s(D)$ とそれぞれ等価である。

ここで、文書集合DではRより少ないB個の文書が話題を共通にしており、他は互いに関連のないものと仮定する。このとき、Aの値がBと等しいか小さければB個の文書に現れる用語の寄与によって行列 T^A において値が0でない成分の和は大きくなり、不一致許容形の文書集合共通度も大きくなる筈である。一方、Aが $R \sim B+1$ の間にあるときは、偶発的にB個以上の文書で現れる用語があったにしてもその用語の各文書内の頻度は高くないものと想定され、 T^A での0でない成分の和は小さく、不一致許容形の文書集合共通度も小さいものと考えられる。従って、ブロック49では求められた不一致許容形の文書集合共通度と予め決められた閾値との比較を行い、閾値以上のときには、行列 T^A は話題を共有する文書の影響を受けている可能性が高いと判断してブロック50に進む。閾値よりも小さければブロック52に進む。この場合の閾値は実験的に決めておく。

【0042】

ブロック50では、行列 T^A を用いて各文書に対して不一致許容形の文書・文書集合共通度を算出し、その値が一定値を越える文書を選択する。 T^A を用いる不一致許容形の文書・文書集合共通度は式(7)式(8)式(9)式(10)において、 S^C_{mn} を T^A_{mn} により置き換えることにより得ることができる。例えば、式(9)式(10)を用いる場合、文書 D_r に対する文書・文書集合共通度を線形モデルでは $com_1(D, D_r; T^A)$ 、2次モデルでは $com_s(D, D_r; T^A)$ と表記すると、これらは以下

のように求めることができる。

【 0 0 4 3 】

【数 1 8】

$$com_1(D, D_r; T^A) = \left[\frac{\sum_{m=1}^M T^A_{mm} S^r_{mm}}{\sqrt{\sum_{m=1}^M (T^A_{mm})^2} \sqrt{\sum_{m=1}^M (S^r_{mm})^2}} \right]^{1/(R-1)}$$

・ ・ ・ (1 3)

【 0 0 4 4 】

【数 1 9】

$$com_s(D, D_r; T^A) = \left[\frac{\sum_{m=1}^M \sum_{n=1}^M T^A_{mn} S^r_{mn}}{\sqrt{\sum_{m=1}^M \sum_{n=1}^M (T^A_{mn})^2} \sqrt{\sum_{m=1}^M \sum_{n=1}^M (S^r_{mn})^2}} \right]^{1/(R-1)}$$

・ ・ ・ (1 4)

【 0 0 4 5 】

文書選択では、線形モデルを採用する場合には $com_1(D, D_r; T^A)$ が、2次モデルを採用する場合は $com_s(D, D_r; T^A)$ が予め設定された閾値を越える文書を選択する。閾値は実験的に決めておく。

ブロック51では、ブロック50において選択された文書集合の文書集合共通度を算出し、文書集合として話題が共通しているか否かを判断するために閾値処理を行う。選択された文書集合の文書集合共通度は線形モデルを採用する場合には式(4)、2次モデルを採用する場合は式(6)を用いて求めることができる。文書集合共通度が閾値以上の場合、またはA=1の場合には処理を終了し、閾値未満の場合には、ブロック52において不一致許容閾値を小さくなるように変更してブロック49に戻り、処理を続行する。

【 0 0 4 6 】

【発明の効果】

ここで本発明の効果を説明する為に図4の実施例に沿った実験結果を示す。実験に用いたデータは文書分類用コーパスReuters-21578から取り出した21記事であり、話題によって3グループに分けられる。内容は、

グループ1：カテゴリ"acquisition"から取り出したGenCorp社の企業買収に関する12記事、

グループ2：カテゴリ"crude"から取り出したエクアドルの地震に関する6記事、

グループ3：カテゴリ"money-fx"から取り出したJames Baker氏の発言に関する3記事、

である。

【0047】

この実験の目的は、21記事の中から文書数の最も多いグループ1を抽出し、さらにグループ1の共通話題を最もよく表す3つの文を選択することである。文数は250で、全用語数は1147であった。ブロック47での共起計数行列Uの各成分の中で最大の値は12でなく、13であった。これは特定の用語がグループ1とグループ2の両方の文書に現れたためである。そこで、Aの初期値を13として図4の49→50→51→52→49の繰り返し処理を行った。ブロック49で得られた不一致許容形の文書集合共通度は、A=13の場合線形モデル、2次モデルとも0.22、A=12の場合には同じく0.39であった。この場合、最も文書数の多いグループ1は文書数が12なので、A=12の不一致許容形の文書集合共通度の方が値が大きいことが期待されたが、結果は期待に沿うものであった。しかし、A=13の場合も値は十分に小さいとは言えないので、A=13の場合もブロック50に進むとした。ブロック50では文書選択のための閾値を線形モデル、2次モデルとも0.02と設定すると、A=13の場合には13文書（グループ1の9文書とグループ2の3文書）、A=12の場合には12文書（全てグループ1）が選択された。選択された文書に対して文書集合共通度を求めると、A=13の場合線形モデルで0.29、2次モデルで0.33、A=12の場合にはそれぞれ0.85、0.90と得られた。従って、ブロック51における閾値が0.5となっていれば、A=12のときに選択された12文書が話題の揃った文書として出力されることになるが、前述のようにこれらは全てグループ1に属している。

【0048】

また、選択された文書に対し、式（9）を用いて各文の文・文書集合共通度を求め、値の大きな文を3個選択した結果を以下に示す。

1位：General Partners said it is asking GenCorp for its shareholder list

s for help in disseminating the offer.

2位: Earlier today, General Partners, owned by Wagner and Brown and AFG Industries Inc, launched a 100 dlr per share tender offer for GenCorp.

3位: General Acquisition Co said it was disappointed by Gencorp's response to its tender offer and asked how the company might give better value to shareholders.

これにより、文書集合で最も優勢な話題はGenCorp社の企業吸収に関するものであることが分かる。また、文書集合から選択された文書を除去して同様な処理を行えば2番目に優勢な話題を述べた文書（この場合にはグループ2）を抽出することができる。

【 0 0 4 9 】

このように本発明によれば、文書集合中で最も優勢な話題を共有する文書群を取り出し、同時に共通話題に最も近い文をユーザに提示することができる。そのためユーザの情報取得の効率性が高められる。

【図面の簡単な説明】

【図 1】

本発明の概略を示すブロック図である。

【図 2】

本発明の不一致許容形共通共起行列をの作成方法を示す図である。

【図 3】

文書集合が入力された段階から文書集合共通度、文書（文書セグメント）・文書集合共通度が決定されるまでの手順を示す図である。

【図 4】

文書集合が入力された段階から最も優勢な話題を述べた文書が抽出されるまでの手順を示す図である。

【符号の説明】

1 1 0 : 文書入力ブロック

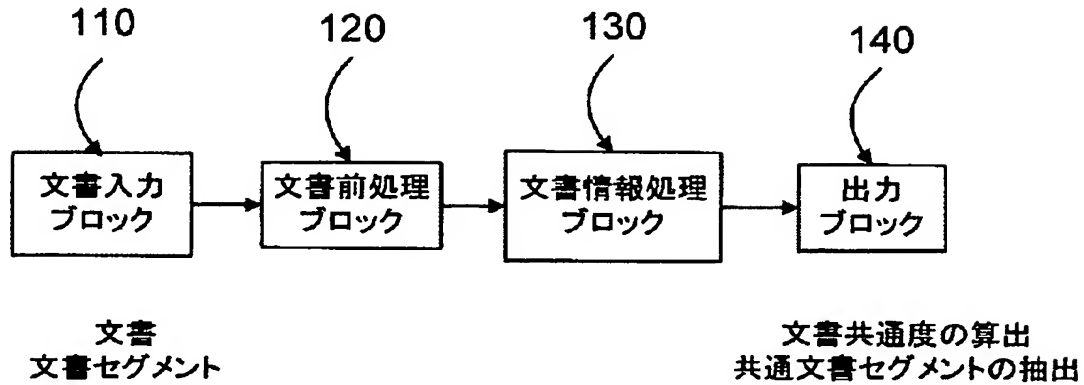
1 2 0 : 文書前処理ブロック

1 3 0 : 文書情報処理ブロック

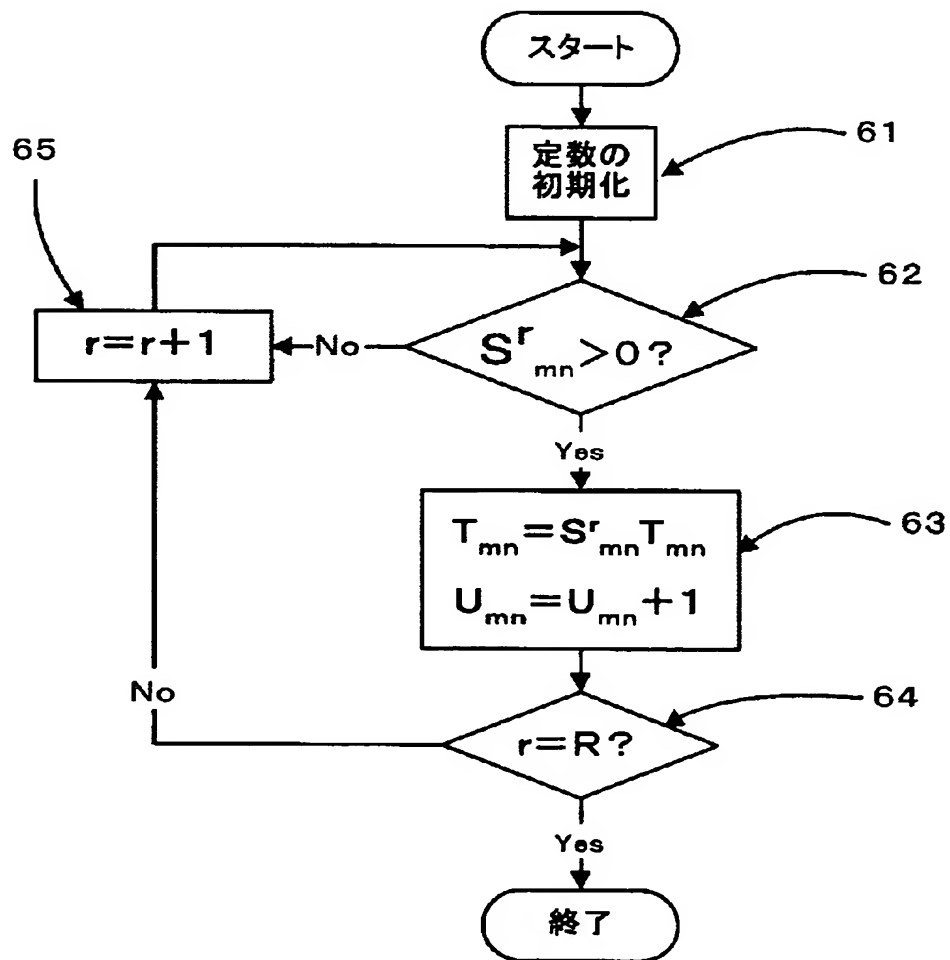
1 4 0 : 出カブロック

【書類名】 図面

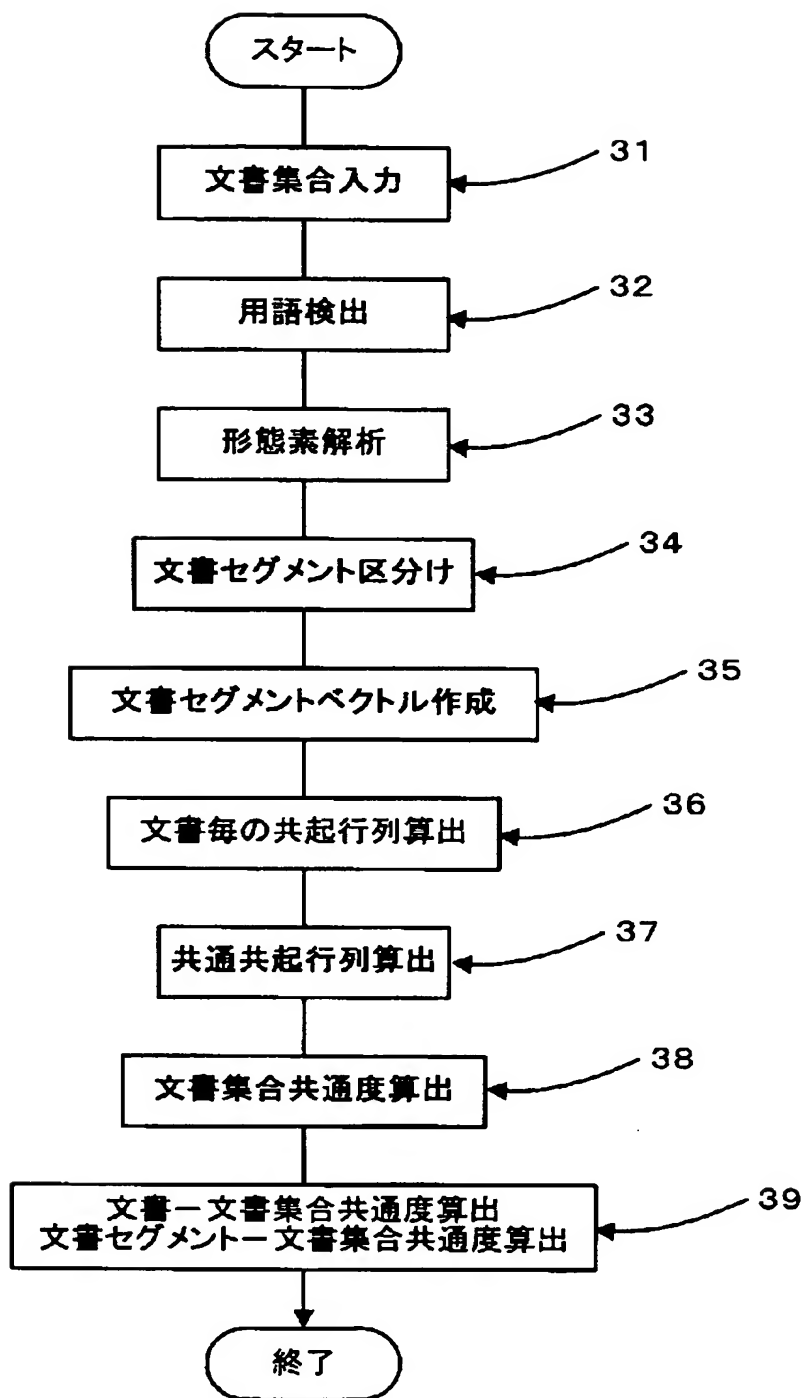
【図 1】



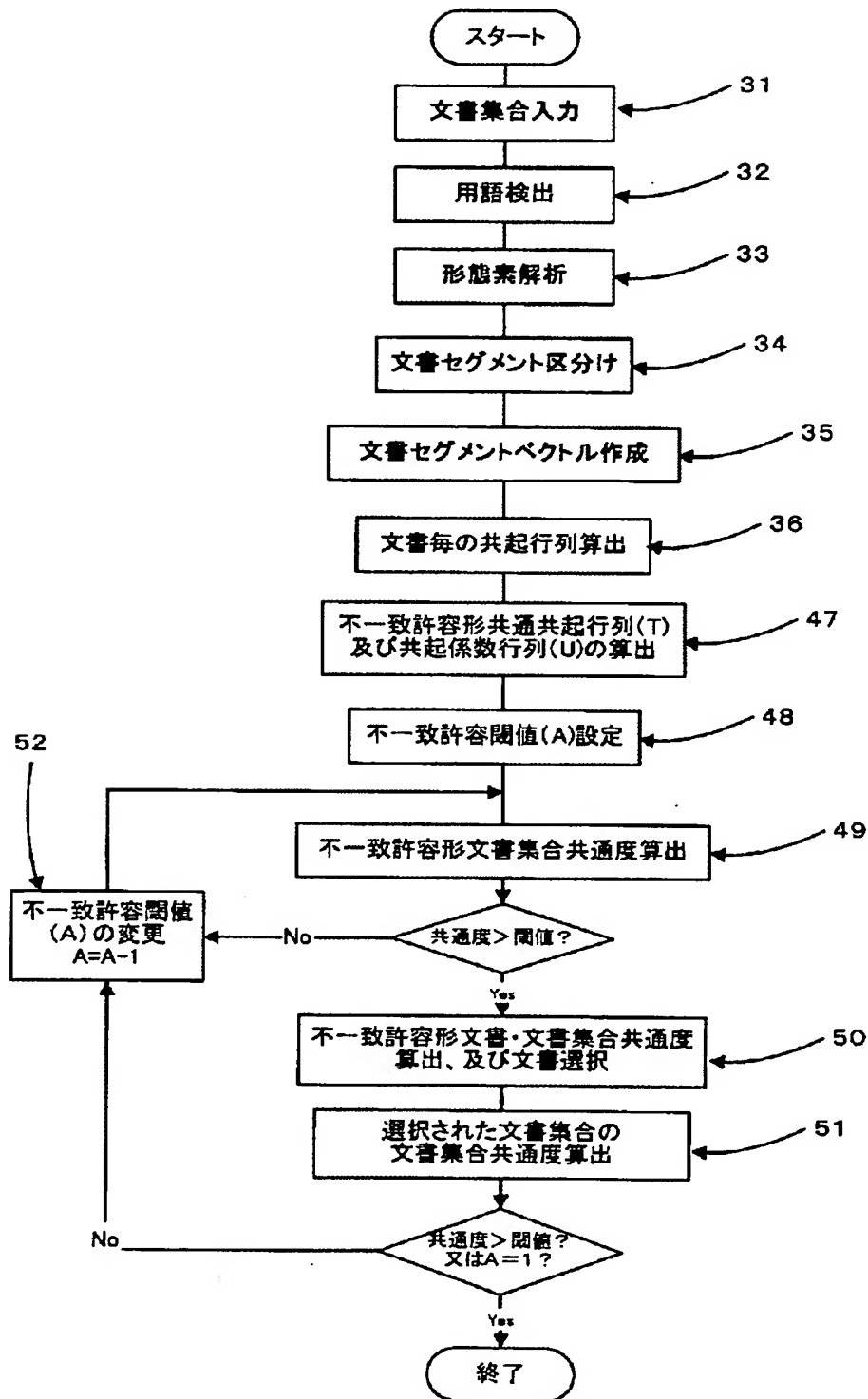
【図 2】



【図 3】



【図 4】



【書類名】 要約書

【課題】

自然言語処理において3個以上の文書がどの程度話題を共通にしているかを表す尺度はこれまで知られていなかった。また、必ずしも話題が同じでない文書集合からの共通の話題を述べた文書の抽出、各文書、各文への共通話題への近さに応じたスコア - の付与は、従来のクラスタリング技術では完全ではなかった。

【解決手段】

本発明では、各文を各成分が対応する用語の有無を表す2値ベクトルで表したうえで、文書間の共通ベクトルの概念を導入する。共通ベクトルは、各文書から1つずつ取り出した文ベクトル群において全てのベクトルで1となる成分のみが1となり他はゼロとなるようなベクトルである。各共通ベクトルにおける値が非ゼロの成分数の全共通ベクトルに対する和、もしくは2乗和を用いることにより、文書集合の共通度を求めることができる。また、各文を全共通ベクトルに射影し、射影値の和等により、各文が共通話題にどの程度近いかを知ることができる。

図 3

認定・付加情報

特許出願の番号	特願 2 0 0 2 - 3 2 6 1 5 7
受付番号	5 0 2 0 1 6 9 4 3 1 6
書類名	特許願
担当官	第七担当上席 0 0 9 6
作成日	平成 1 4 年 1 1 月 1 1 日

< 認定情報・付加情報 >

【提出日】 平成14年11月 8日

出 願 人 履 歴 情 報

識別番号 [398038580]

1. 変更年月日 1998年 5月19日

[変更理由] 新規登録

住 所 アメリカ合衆国カリフォルニア州パロアルト ハノーバー・ストリート 3000

氏 名 ヒューレット・パッカード・カンパニー